The Rasch Model for Achievement Tests–Inappropriate in the Past, Inappropriate Today, Inappropriate Tomorrow

Leslie D. McLean & Ronald G. Ragsdale
the ontario institute for studies in education

Robitaille and O'Shea (1983) used the Rasch model in developing a set of 40-item mathematics achievement tests using multi-choice items for the British Columbia item bank. The present paper argues that (a) the Rasch model is inappropriate for such a purpose, (b) their article presents the model in a far more favourable light than it deserves, and (c) continued application of the model encourages less desirable educational practice and misdirects scarce resources. A fundamentally different approach is suggested and illustrated by work on the Ontario Assessment Instrument Pool (OAIP).

Lors du développement de tests de rendement en mathématiques comprenant une série de 40 items à choix multiples, Robitaille et O'Shea se sont basés sur le modèle Rasch afin de former la banque d'items de la Colombie Britannique. Le présent travail soutient que: (a) le modèle Rasch répond inadéquatement à un tel but; (b) leur article surestime le modèle; (c) l'application continue de ce modèle encourage l'emploi de mesures éducationnelles médiocres et une mauvaise gestion des ressources. Une approche essentiellement différente est alors suggérée et démontrée à partir du travail qui a été fait auprès de la banque d'instruments de mesure de l'Ontario (BIMO).

Robitaille and O'Shea list "several advantageous properties of the [Rasch] model." These are (1) "within reasonable limits the estimates of item difficulty are independent of the sample of persons used to calibrate the items," (2) "within reasonable limits the abilities of two persons may be assessed using completely different items, and the estimates of their abilities will be directly comparable," and (3) "the best estimate of a person's ability is obtained from a test consisting of items each of which has a probability of success for that individual of 0.5."

Several comments are offered about the first "advantage." First, much depends on the meaning of the weasel words "within reasonable limits." If samples come from the same population, traditional statistical theory gives us reasonable limits within which the estimates would be regarded as equivalent. The rules for the reasonable in the Rasch framework are not specified. Second, we must note the later statement by Robitaille and O'Shea, that "since the booklets were calibrated independently, the difficulty of each item in link *a* assumes two values: one resulting from its inclusion in booklet *A*, and one from its inclusion in booklet *B*." They go on to explain how one can adjust the item difficulties from one booklet to make them correspond to those in the other booklet, leading one to

suspect that the first advantage is *ad hoc* and leaves something to be desired.

The second advantage contains the same disclaimer, "within reasonable limits," leading to the same uncertainty. If the words mean that the *items* are drawn from the same domain, the classical test theory notion of equivalent forms would seem to lead to an agreement in conclusions. Fitting all types of items to the same scale, however, implies that geometry ability can be tested by items on consumer mathematics, a concept that contradicts both curriculum and test design. Why do the authors retain the categories of algebra and the like, given "advantage" 2? Since the outcome of the B.C. process was a different 40-item test for each content-by-grade-level combination, it appears that the second advantage was not an important factor in creating the tests.

The third "advantage" regarding the desirability of items with a difficulty level of 0.5 is puzzling, since it is a part of classical test theory that has been known and used for many years. When the tests are used with a wide and diverse sample of students, however, both the classical result and the Rasch rediscovery are simply not credible. The same test cannot simultaneously provide the best measure of ability of high and low ability students. The following *may* be true: *if* you can only have one test, *then* the best test overall is one with all items at 0.5 difficulty.

But why construct only one test (or even a few 40-item tests) for every student in British Columbia? Apart from prize competitions (for which one would choose items with difficulties about 0.8), there appears no good reason to give everyone the same test. Given the large item pool, why not try for varying length tests built by each teacher for the low, average, and high ability students? Using the items in this way will be more work for teachers, but the teaching and diagnostic value would be great. Primary interest will not concern the total number of correct responses (a test score), but rather the items which were answered correctly. The classical concern for reliability of a test score is replaced by a concern for the validity of the content. In other words, we argue that a better response to the "increased demand for valid and reliable achievement tests keyed to the prescribed curriculum for use by classroom teachers" (Robitaille & O'Shea, p. 57) would be a response that makes use of the riches of the item pool rather than one that harks back to the time when the lack of one made it necessary to rely on fixed tests that fitted no teacher's curriculum well.

In total then, the three advantages of Rasch scaling sound attractive, but they don't hold up. Moreover, they bear little relation to the B.C. product. We agree with Goldstein and Blinkhorn (1977) and Traub and Wolfe (1981) that the Rasch model of responses is inappropriate for these multi-choice achievement test items. Virtually all the necessary assumptions are untenable. A single, homogeneous, latent trait is assumed, whereas the item set has been carefully constructed to be heterogeneous. Responses are assumed to be independent (more precisely, to have the

characteristics of local independence), whereas inclusion in one 40-item test which all students must answer in the same order almost surely introduces dependencies that make the original calibrations suspect (if they ever were valid). Responses are assumed to depend only on students' ability, whereas four-option multi-choice items are certain to attract some guessing. The above arguments appear plausible to the present authors and Robitaille and O'Shea report no tests of their assumptions. Indeed, so questionable are the tests of assumptions provided by Robitaille and O'Shea's BICAL program, it is fortunate that they were not reported.

Robitaille and O'Shea might give the impression that the Rasch model as implemented via BICAL was a thoroughly tested and proven tool. They acknowledge that there have been criticisms, but cite the Portland Public Schools' experience and a 1978 ERIC reference (Rentz & Rentz, 1978) as evidence that the Rasch model is a viable alternative. In contrast to the Portland Public Schools, however, the National Foundation for Educational Research in England and Wales (NFER) has abandoned Rasch scaling after years of experience; the U.S. national assessment has *never* used it. The NFER decision went much farther than "a reconsideration of the appropriateness of using the Rasch model as a means of monitoring changes in patterns of educational achievement over time" (Robitaille and O'Shea, p. 60). The problems with Rasch scaling are not just with measurement of change over time, and it is still possible to find a definitive answer to questions raised.

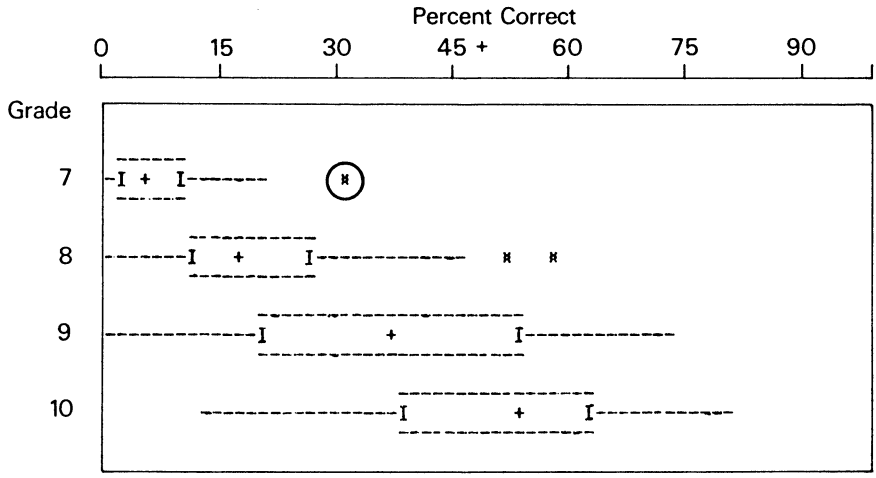
THE OAIP ALTERNATIVE

If, as we argue, Rasch scaling is inappropriate, how can teachers be effectively presented with information concerning the test items created for their use? How do we propose to make the items useful for teachers? Space limitations prevent a complete description, but we can outline our alternative with some aspects of our recent work on the Ontario Assessment Instrument Pool (OAIP).

While OAIP resembles in many ways the item pools in existence elsewhere, it has some features which make it unique. An extremely important feature is that all instruments in each subject area are published and sent to the schools for use by teachers at the end of the development phase. The broad term "instrument" is used rather than "item" because OAIP instruments are not restricted to multi-choice items. Many of them, including virtually all the mathematics items call for a constructed response. Like other pools (B.C. included) the instruments are closely linked to the provincial curriculum guidelines.

The provision of many instruments keyed to objectives in the guidelines eliminates the need for fixed tests – hence any need to scale and equate them. Measurement of achievement can be tailored locally to local needs, as the evidence suggests it ought to be.

FIGURE 1
Range of class means on the 24 algebra instruments, by grade



Statistical Summary

Grade Level	No. of Classes	Grade Mean	Grade St. Dev.
7	81	6.6	6.0
8	89	19.1	11.6
9	92	36.8	19.9
10	74	49.8	17.2

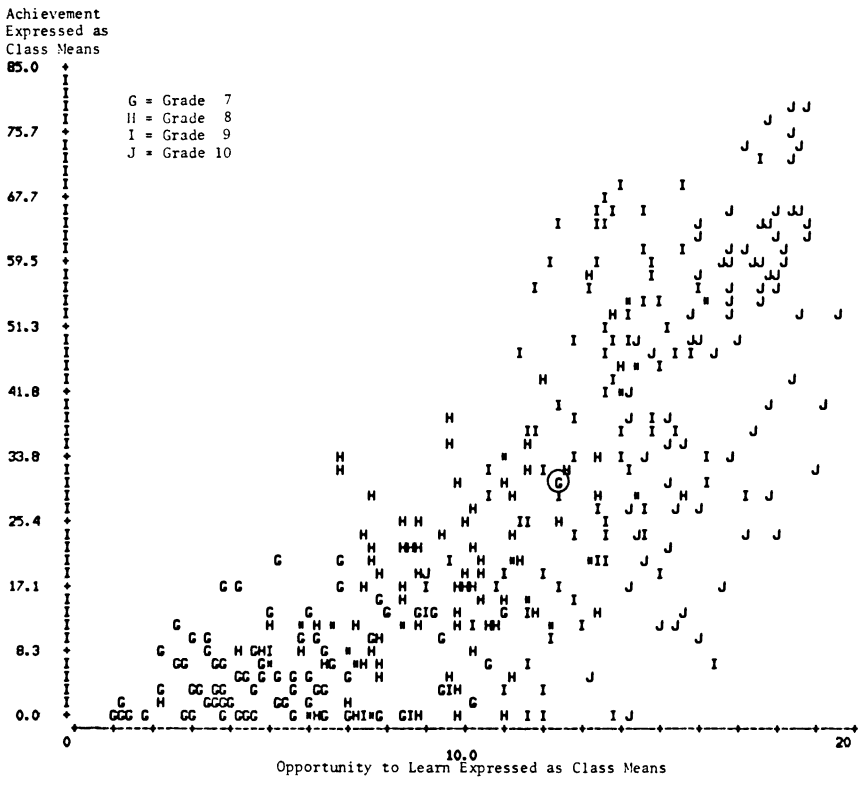
Consider this one result from the 1981 province-wide administration of OAIP mathematics instruments in Grades 7, 8, 9, and 10 (McLean, 1982). Four algebra instrument forms from the pool were used to generate 24 explicit instruments (six each) among a total of over 600 instruments. The 24 were distributed randomly among 18 booklets of 30 instruments each. Booklets were distributed randomly to students in 366 classes. For each instrument, students were asked both to write in their answer and also to check when they were taught the information needed to answer the question – before, this year, or not yet. This information is referred to as Opportunity To Learn (OTL).

The ranges of class means on the 24 instruments (percentage correct) are presented in box and whisker plots in Figure 1. The boxes span the middle 50% of class means. The asterisk circled in Grade 7 marks a class whose mean is 1.5 or more box lengths above the box (an outlier).

OTL class means are plotted against achievement class means in Figure 2. The pooled *within grade* correlations between OTL and achievement is 0.52. From other information, we were able to verify that the student OTL responses agreed very well with the teacher responses indicating

FIGURE 2

Class means on the 24 algebra instruments versus Opportunity To Learn, for each of four grades



when the topic is taught. The outlying Grade 7 class (the circled "G" in the centre) is easily explained. Students reported overwhelmingly that they had been taught that material "this year," in marked contrast to the other Grade 7 classes.

For any OTL value (taught this year, for example), a wide range of achievement was observed. This is not surprising, since classes differ in ability, amount of emphasis on topic, and in many other ways. The best way to determine whether students have learned what they should have learned in their classroom or school is to tailor measurement to local needs. Instrument pools make it feasible to tailor measurement to local needs. Why would we settle for set tests (prize competitions excepted)?

The Rasch model is now seen as an inappropriate distraction, a piece of fools' gold. It looks good at first, but turns out on close inspection to be a worthless waste of time.

REFERENCES

Goldstein, H., & Blinkhorn, S. Monitoring educational standards – An inappropriate model. *Bulletin of the British Psychological Society*, 1977, *30*, 309–311.

McLean, L.D. *Report of the 1981 field trials of the Ontario Assessment Instrument Pool.* Toronto: Ontario Ministry of Education, 1982.

Rentz, R.R., & Rentz, C.C. *Does the Rasch model really work? A discussion for practitioners.* Princeton: ERIC Clearinghouse on Tests, Measurement, and Evaluation, E.T.S., 1978.

Robitaille, D.F., & O'Shea, T. The development of an item bank in mathematics using the Rasch model. *Canadian Journal of Education*, 1983, *8*(1), 57–70.

Traub, R.E., & Wolfe, R.G. Latent trait theories and the assessment of educational achievement. In D.C. Berliner (Ed.), *Review of Research in Education*, 1981, *9*, 377–435.

Leslie D. McLean is a Professor and Ronald G. Ragsdale is an Associate Professor in the Department of Measurement, Evaluation, & Computer Applications, The Ontario Institute for Studies in Education, 252 Bloor Street West, Toronto, Ontario M5S 1V6.